

CLAIMS

1. A system for summarising data sets, the system comprising:
a target data item store for storing target data items;
5 sectioning means for dividing said data set into sections and for comparing
each section against said target data items;
calculation means for calculating a ranking value for each said section,
said ranking value dependent on the outcome of said comparisons; and
compilation means for compiling a summary of the data set by selecting
10 one or more section(s) according to the respective ranking values.
2. A system according to claim 1, further comprising a user input for
inputting target data items to the target data item store.
- 15 3. A system as claimed in either one of the preceding claims, the system
further comprising:
a key data item identifier for identifying key data items of said data set;
a distribution value calculator for calculating a distribution value for each
section dependent on the distribution of said key data items in said section; and
20 ranking value adjustment means for adjusting the relevant ranking value in
a manner dependent on said distribution value for each section.
4. A system as claimed in any one of claims 1 to 3 wherein said sections
within a compiled summary are ordered according to the order of their occurrence
25 in the data set.
5. A system as claimed in any one of claims 1 to 3 wherein said sections
within a compiled summary are ordered according to their ranking value.
- 30 6. A system as claimed in any one of claims 2 to 5 wherein said distribution
value calculator calculates said distribution value for each section by:
determining a first score for each key data item in each section; and
for each section, summing said first scores for each key data item,

wherein said first score of each key data item is calculated as the number of times the key data item of consideration occurs in the data set less the number of times the key data item of consideration occurs in the section of consideration.

- 5 7. A system as claimed in any one of claims 2 to 6 wherein said distribution value calculator calculates or modifies said distribution value for each section by calculating a second score for each key data item, said second scores being calculated by:

assigning a position value to each section of the data set corresponding to the
10 position of the section within the data set; and
for each key data item of the data set, performing the calculation of subtracting the position value of the first section in which the key data item of consideration occurs from the position value of the final section in which the key data item of consideration occurs.

15 .

8. A system as claimed in claim 7 wherein said distribution value calculator calculates or modifies said distribution value for each section by calculating a third score for each key data item and summing the third scores calculated for each
20 section, said third score being calculated by:

assigning a position value to each section of the data set corresponding to the position of the section within the data set;
identifying every pair of sections in which key data items co-occur;
for each identified pair of sections, subtracting the lower position value from the
25 higher position value and dividing the result by the second score.

9. A method for summarising data sets comprising the steps of :

- 30 1) receiving a data set as input to processing means;
2) storing one or more target data items in a target data store;
3) dividing said data set into sections;

4) comparing each said section against said target data items;

5) calculating a ranking value for each said section dependent on the outcome of said comparison; and

5

6) compiling a summary of the data set by selecting one or more section(s) according to the respective ranking values.

10. A method according to claim 9 further comprising the step of receiving
10 said one or more target data items at a user input.

11. A method as claimed in either one of claims 9 or 10 further comprising the steps of :

15 7) identifying key data items within said data set;

8) calculating a distribution value for each said section dependent on the distribution of said key data items; and

20 9) modifying said ranking value dependent on said distribution value.

12. A method as claimed in any one of claims 9 to 11 wherein the sections within a compiled summary are ordered according to the order of their occurrence in the data set.

25

13. A method as claimed in any one of claims 9 to 11 wherein the sections within a compiled summary are ordered according to the ranking value of step 5.

14. A method as claimed in any one of claims 11 to 13 wherein said
30 distribution value is calculated or modified for each section by:
determining a first score for each key data item in each section; and
for each section, summing said first scores for each key data item,

wherein said first score of each key data item is calculated as the number of times the key data item of consideration occurs in the data set less the number of times the key data item of consideration occurs in the section of consideration.

5 15. A method as claimed in any one of claims 11 to 14 wherein said distribution value is calculated or modified for each section by calculating a second score for each key data item, said second scores being calculated by assigning a position value to each section of the data set corresponding to the position of the section within the data set and, for each key data item of the data set, performing
10 the calculation of subtracting the position value of the first section in which the key data item of consideration occurs from the position value of the final section in which the key data item of consideration occurs.

16. A method as claimed in claim 15 wherein said distribution value is
15 calculated or modified for each section by calculating a third score for each key data item by identifying every pair of sections in which key data items co-occur, for each pair of sections subtracting the lower distribution value from the higher distribution value and dividing the result by the second score, summing the third scores calculated for each section whereby calculating a fourth value for each
20 section, and calculating or modifying the distribution value in accordance with said fourth value for each section.

17. Apparatus for summarising data sets, the apparatus comprising:
i) a data set input for receiving a data set;
25 ii) means for dividing a received data set into sections; and
iii) means for processing the sections so as to output a summary of the data set which comprises selected sections thereof,

wherein the means for processing the data set comprises either one or both of:

iv) an input for receiving at least one set of target data items; and
30 v) means for generating a set of key data items from the received data set received at the data set input,

and wherein the means for processing the data set is adapted to select sections for use in a summary by means of detecting target data items therein and/or by means of detecting a distribution of key data items in the data set.

18. Apparatus according to claim 17, wherein the apparatus is adapted to select sections for use in a summary by means of first detecting target data items therein, assigning a ranking value to each section in accordance with the presence
5 or otherwise of target data items in the section, and modifying the ranking value in accordance with a detected distribution of key data items in the data set.